

## ARTICLE

# A decision support system based on artificial intelligence and systems biology for the simulation of pancreatic cancer patient status

Valentin Junet<sup>1,2</sup>  | Pedro Matos-Filipe<sup>1,3</sup>  | Juan Manuel García-Illarramendi<sup>1,2</sup>  |  
 Esther Ramírez<sup>4</sup> | Baldo Oliva<sup>3</sup>  | Judith Farrés<sup>1</sup>  | Xavier Daura<sup>2,5,6</sup>  |  
 José Manuel Mas<sup>1</sup> | Rafael Morales<sup>7</sup>

<sup>1</sup>Anaxomics Biotech SL, Barcelona, Spain

<sup>2</sup>Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain

<sup>3</sup>Structural Bioinformatics (GRIB-IMIM), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain

<sup>4</sup>Iteraset Solutions SL, Barcelona, Spain

<sup>5</sup>Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

<sup>6</sup>Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Instituto de Salud Carlos III, Cerdanyola del Vallès, Spain

<sup>7</sup>Hospital La Mancha Centro, Alcázar de San Juan, Spain

## Correspondence

Judith Farrés, Anaxomics Biotech SL, 08007 Barcelona, Spain.

Email: [judith@anaxomics.com](mailto:judith@anaxomics.com)

Xavier Daura, Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain.

Email: [xavier.daura@uab.cat](mailto:xavier.daura@uab.cat)

## Abstract

Oncology treatments require continuous individual adjustment based on the measurement of multiple clinical parameters. Prediction tools exploiting the patterns present in the clinical data could be used to assist decision making and ease the burden associated to the interpretation of all these parameters. The goal of this study was to predict the evolution of patients with pancreatic cancer at their next visit using information routinely recorded in health records, providing a decision-support system for clinicians. We selected hematological variables as the visit's clinical outcomes, under the assumption that they can be predictive of the evolution of the patient. Multivariate models based on regression trees were generated to predict next-visit values for each of the clinical outcomes selected, based on the longitudinal clinical data as well as on molecular data sets streaming from in silico simulations of individual patient status at each visit. The models predict, with a mean prediction score (balanced accuracy) of 0.79, the evolution trends of eosinophils, leukocytes, monocytes, and platelets. Time span between visits and neutropenia were among the most common factors contributing to the predicted evolution. The inclusion of molecular variables from the systems-biology in silico simulations provided a molecular background for the observed variations in the selected outcome variables, mostly in relation to the regulation of hematopoiesis. In spite of its limitations, this study serves as a proof of concept for the application of next-visit prediction tools in real-world settings, even when available data sets are small.

## Study Highlights

### WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Artificial intelligence (AI) and related technologies are increasingly prevalent in health care, as they can manage heterogeneous sources of data, identifying underlying patterns, and predicting complex outcomes.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *CPT: Pharmacometrics & Systems Pharmacology* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

### WHAT QUESTION DID THIS STUDY ADDRESS?

Evolution of certain patient parameters can be forecasted using AI and assist in the complex assessment of oncology patients under chemotherapy treatment.

### WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

This study serves as a proof of concept for the application of AI to the prediction of next-visit outcomes in real-world settings, highlighting its usefulness even with reduced data sets. The approach presented could forecast the trend of hematological parameters at the next visit with time span between visits and neutropenia as the most common factors involved in the forecast.

### HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

Models, such as the one used in this study, where the decision process can be mapped and interpreted, have the potential to gain the trust of clinicians and medical researchers and become a routine tool in daily clinical practice to guide patient treatments.

## INTRODUCTION

Cancer is a clinical term encompassing several diseases with well-differentiated histologic characteristics, heterogeneous clinical performances, and failed clinical response in many cases. Pancreatic cancer is one of the most aggressive cancer types, with a 5-year survival rate of less than 5%.<sup>1</sup> A lack of early detection methods along with a tendency for early metastasis contributes to this poor survival rate.<sup>2</sup> Minor improvements have been obtained with folinic acid, 5-fluorouracil [5-FU], irinotecan, and oxaliplatin (FOLFIRINOX) and paclitaxel/nab-paclitaxel plus gemcitabine chemotherapy.<sup>3,4</sup>

Patients with cancer, especially those in advanced stages, usually display a nonrecurrent clinical behavior. Oncologic treatments may lose effectiveness over time due to the appearance of resistance. Besides, oncologic treatments may have toxic effects, requiring dose adjustment, therapy discontinuation, or substitution. Additional treatments may be prescribed to reduce oncologic treatment toxicity and for comorbid illness management. Thus, a continuous evaluation of patient status and of potential treatment adjustments is necessary.<sup>5,6</sup> Different sets of measurements are used to assess the impact of the tumor on the affected organ, treatment efficacy, and patient's well-being. Blood biomarkers along with resonance-based imaging techniques for solid tumors are used to assess efficacy. Different analytical parameters are measured in blood to evaluate liver, kidney, and immune-system function. Patient's quality of life and ability to perform daily routine tasks are measured using performance scales like Eastern Cooperative Oncology Group (ECOG) and Karnofsky Performance Scale Index. Evaluating all this patient and treatment information for decision making is not an easy procedure, and

the oncologist would greatly benefit from reliable forecast systems based on current patient data, predicting the potential effect the treatment will have on the follow-up tests.

Computational tools based on mathematical models for medical image analysis have been the earliest to be applied to assist the grading of a disease.<sup>7-9</sup> Lately, we have seen a surge of proof-of-concept uses of artificial intelligence (AI) in medicine, due to both the increase in data availability and the advances in AI algorithms, which allow their use under real-life conditions to assist human decision makers.<sup>10</sup> In order to incorporate molecular profiles in clinical decision-making processes, many approaches combine AI with systems-biology strategies interpreting the interactions between extensive molecular measurements.<sup>11,12</sup> One such approach is the Therapeutic Performance Mapping System (TPMS).<sup>13</sup> The modeling of TPMS encompasses all known relationships in the human proteome and information on drugs and diseases found in accessible databases. This wide data spectrum makes it possible to derive a patient-specific pattern of protein activity based on the clinical status (main disease, comorbidities, and adverse events) and treatments taken by the patient. This approach has been used to study numerous pathologies, including several cancer types.<sup>14-16</sup>

The objective of the present study was to perform computational simulations to predict the evolution of clinical variables in patients with metastatic pancreatic cancer at their next visit and evaluate this predictor as a system to support clinical decisions during the treatment of patients with pancreatic cancer. Model generation was done using supervised machine-learning techniques on data from real patients enrolled in the SICPAC study, who underwent first-line palliative chemotherapy under the regime gemcitabine/nab-paclitaxel. The concentrations of hemoglobin,

red blood cells, eosinophils, leukocytes, monocytes, and platelets were selected as clinical-outcome variables for the analysis. Other goals of the study included the identification of those patient characteristics, treatment options, and time intervals between cycles that might influence the evolution of the selected outcome variables. The inclusion of data from systems biology was expected to highlight proteins and pathways that might have a key role in the variations observed in these outcomes.

## METHODS

### SICPAC study data

The SICPAC study is an observational study with authorization prior to recruitment carried out at La Mancha Centro Hospital, Alcázar de San Juan, Ciudad Real, Spain, devoted to the retrospective monitoring of patients with pancreatic cancer. It is classified as an observational post-authorization study by the Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), with registration number MOR-GEM-2018-01.

The study enrolled 20 patients diagnosed with locally advanced or metastatic adenomatous pancreatic cancer from 2015 to 2018. Patients had previously undergone palliative first-line chemotherapy with gemcitabine/nab-paclitaxel and were over 18 years old at the time of recruitment. The availability of clinical information of at least three visits (1 treatment cycle) of each patient was also required.

Patient data referring to demographics, previous personal records, symptoms of disease, pharmacologic treatment, physical exploration, and tumor biology was retrieved from the Case Report Form (CRF). The study collected a total of 40 clinical variables listed in [Table S1](#).

### Simulation of patients by systems biology

Data collected from the SICPAC study was used to model individually the 20 recruited patients using the TPMS systems-biology approach. The methodology used has been described in detail elsewhere.<sup>13,17</sup> Briefly, mathematical models are built on the basis of a human-protein functional network that incorporates the available relationships (edges or links) between proteins (nodes) from a regularly updated in-house database drawn from public sources. A selected collection of known input-output physiological signals (such as drug-indication pairs) are collated to train the models. The information relating biological processes (adverse drug reactions, indications, and diseases) to their molecular effectors (i.e., proteins described to be involved

in the pathophysiological process), is compiled in the biological effectors database (BED), currently describing more than 300 clinical conditions. The TPMS algorithm takes as input signals the activation (+1) and inactivation (−1) of the drug's target proteins, and as output the BED protein states of the pathology. It then optimizes the paths between both protein sets and computes the activation and inactivation values of all proteins in the network. Each node of the protein network receives as input the output of the incoming connected nodes and every link is given a weight ( $\omega$ ). The sum of inputs is transformed by a hyperbolic tangent function that generates a score for every node, which becomes the “output signal” toward the outgoing connected nodes. The  $\omega$  parameters are obtained by optimization, using a Stochastic Optimization Method based on Simulated Annealing. Because the number of entries in the training set is always smaller than the number of parameters (link weights) required by the algorithm, a population of solutions with accuracies over 95% are obtained.

In this study, a different TPMS model is built per patient and visit, making 274 different models in total. The drugs taken by a patient (treatments and co-treatments) at each visit are considered inputs for the model, in the form of activation or inhibition of their corresponding protein targets. Gender and anthropometric measures have been used to adjust the signal received by the drug targets. The target population for the analysis is the protein set describing pancreatic cancer as described in BED. Adverse events or comorbidities included in the CRF are also incorporated in the form of activation/inhibition of the protein sets describing the conditions in BED. We have calculated over 250 solutions per model that closely reproduce the clinical information of each patient and visit described in the CRF. The output of each model is the activation/inhibition pattern of a set of proteins and biological pathways.

### Artificial intelligence analysis

Time series analysis was performed using features identified in the CRF, both numeric and categorical, as well as the protein pattern obtained with TPMS. Multivariate models based on regression trees were generated to predict next-visit values of the six clinical outcomes selected (concentration of eosinophils, platelets, red blood cells, leukocytes, monocytes, and hemoglobin).

All consecutive patient visits noted in the CRF were used to train the models, where two visits are considered as successive if the time period between them is less than 50 days. Categorical clinical variables were first split per category (e.g., the CRF clinical variable “co-treatment” was split into as many CRF model variables as different

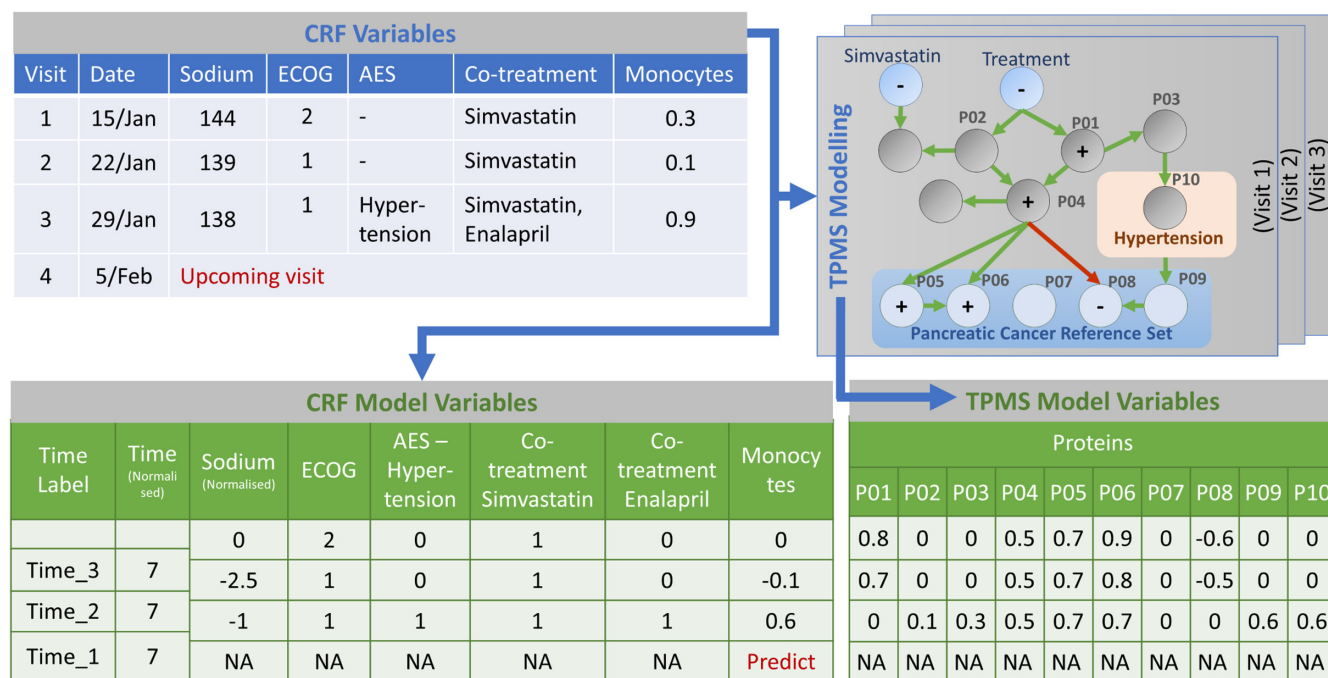
co-treatments were listed) and then labeled as 1 and 0 depending on whether the patient was under the category effect or not, respectively. Continuous variables (including the 6 selected clinical outcomes) were normalized by subtracting from their values the median of all their values up to the last available visit (for each patient separately). The number of days between consecutive visits was also used as a predictive variable and was normalized as such. A representation of the conversion from CRF variables and TPMS variables into model variables is shown in [Figure 1](#).

The model consists of an ensemble of up to five regression trees selecting up to three model variables (4 including the time period between visits) from up to three consecutive visits. For each of the six selected clinical outcomes, a model was trained and its performance was evaluated in a leave-one-patient-out (LOPO) cross-validation setup. In this setup, the data from all visits from all but one patient are used in the training, and all the visits from the left-out patient are used for validation; the training-validation splitting process is repeated for all patients. The number of consecutive visits (1, 2, or 3) used to predict the following one is chosen based on LOPO performance. For example, if we want to predict the outcome at visit 4, and using the variable values from visit 2 and visit 3 leads to better LOPO performance than using the values from visit

1, visit 2, and visit 3, then the chosen number of consecutive visits will be two instead of three.

To evaluate the ability to predict new concentration values at any upcoming visit and detect concentration trends, the Spearman correlation (between the actual values from the CRF and the predicted ones) and a prediction score were used, both with the LOPO setup for cross-validation. To avoid overfitting, the selection of variables was performed within the LOPO setup. Specifically, the feature selection process was performed in nested LOPO setups within the outer/main LOPO setup. These nested LOPO setups are performed with the training data from the outer LOPO setup in two steps. First, each variable is tested individually in a nested LOPO setup and the best performing ones are selected. Second, all possible combinations of up to three variables are tested in nested LOPO setups and the final variables are selected from the best performing combination.

In this study, the protein profiles derived from TPMS models were used as additional variables to the clinical ones, yielding a higher-dimensional training dataset. Thus, for each patient and visit, the TPMS models generate protein activation values that constitute the additional variables. These added variables are treated the same way as the clinical continuous variables, as explained above and in [Figure 1](#).



**FIGURE 1** Scheme of how patient information in the CRF is transformed into CRF and TPMS model variables. CRF variables with information on medical conditions, drug treatments, sex, age, and BMI are used by the TPMS analysis. CRF numerical variables are normalized, categorical variables are transformed to binary for their use in the time series analysis, after conversion of each category in a different model variable. Each of the proteins resulting from the TPMS analysis is used as a model variable. AEs, adverse events; BMI, body mass index; CRF, Case Report Form; ECOG, Eastern Cooperative Oncology Group; NA, not applicable; TPMS, Therapeutic Performance Mapping System.

To better assess the relevance of adding the TPMS protein profiles, two separate analyses, one with the CRF variables and the other with CRF plus TPMS variables, were performed using regression trees.

## Web application

A graphic user-friendly interface has been built and deployed as a webapp, accessible at <http://sicpac.anaxomics.com:81>, to query the models. The tool was built in a python-based environment using the Django back-end.<sup>18</sup> As all the regression trees were built on MATLAB programming language, we used the MATLAB Compiler SDK toolbox as interface between Python and the original code.

## RESULTS

### Patient characteristics

Data analyzed streams from the SICPAC study, an observational clinical study with authorization prior to recruitment devoted to the retrospective monitoring of patients with pancreatic cancer. The study recruited 20 patients diagnosed with locally advanced or metastatic adenomatous pancreatic cancer that had undergone palliative first-line chemotherapy with gemcitabine/nab-paclitaxel and with a follow-up of at least three visits. A total of 274 clinical visits were gathered with a mean time span of 163 days for the whole follow-up. A total of 40 clinical variables per visit were collected consisting of analytical measures, presence of adverse events (AEs), and co-treatments. Out of the 40 clinical variables, 23 were used for the analysis. Variables with many missing values or

with little variation among patients and visits where discarded (Table S1).

The population age mean value at the time of first visit was 64 years (ranging between 48 and 81 years) with a median value of 61 years and a sex ratio of 40:60 women to men. A total of 49 comorbidities and AEs (Table S2) were registered with an average of three comorbidities per visit. A total of 67 different drugs were registered (Table S3) with an average intake of six complementary treatments per visit.

The concentrations of hemoglobin, red blood cells, eosinophils, leukocytes, monocytes, and platelets were selected as outcome variables for the analyses; summary statistics are reported in Table 1.

### Time series analysis

The modeling was done using regression trees, a machine-learning method that can be used for predicting either categories or continuous values based on training data. Regression trees have been proven to be effective methods to handle structured data sets. Multivariate models based on regression trees were generated to predict next-visit values for each of the six clinical outcomes selected. To improve performance, a model was made from an ensemble of trees. To render it interpretable and avoid overfitting, up to three model variables were used per individual tree.

A correlation between the observed (from the CRF) and predicted concentration value could be seen for the six outcome variables (Figure 2).

We evaluated whether extending the number and type of model variables by including molecular details into the time series analysis would increase the accuracy of the predictions and provide a mechanistic understanding of

**TABLE 1** Summary statistics for blood cells and hemoglobin concentration values together with the number of adverse effects and complementary treatments per visit

	All			Female			Male		
	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
Leukocytes, 10 <sup>3</sup> /μL	6.61	3.97	5.8	7.26	4.72	5.95	5.73	2.35	5.7
Monocytes, 10 <sup>3</sup> /μL	0.64	0.50	0.50	0.72	0.57	0.50	0.54	0.37	0.50
Eosinophils, 10 <sup>3</sup> /μL	0.20	0.31	0.10	0.24	0.38	0.10	0.15	0.17	0.10
Platelets, 10 <sup>3</sup> /μL	227.75	142.46	185	234.39	155.25	174	218.63	122.82	192
Red blood cells, 10 <sup>6</sup> /μL	3.85	1.98	3.76	3.79	2.58	3.62	3.94	0.43	3.92
Hemoglobin, g/dL	11.37	1.43	11.2	11.07	1.39	11	11.77	1.4	11.65
N of AEs per visit	2.95	2.18	2	3.04	2.19	3	2.83	2.17	2
N of co-treatment per visit	5.68	4.88	5	5.69	5.07	4	5.65	4.64	5

Note: Mean, median, and standard deviation values are computed for the men, women, and whole population.

Abbreviation: AEs, adverse events.

the proteins and pathways that might play a key role in the variations observed in the outcome variables. To that end, for each patient and visit we evaluated, the activation of the drug targets and the protein sets associated to pancreatic cancer and the comorbidities within a systems-biology-based model (TPMS). The output of each model is the activation/inhibition pattern of a set of proteins and biological pathways. The proteins with different activation signals for each patient and visit were included as additional model variables in the time series analysis. To evaluate the contribution of adding a patient-specific molecular pattern in the analysis, two different sets of regression trees were built to predict each of the six selected outcome variables, one set with only CRF model variables and another set with both CRF and TPMS model variables.

### Ability to predict concentration values of outcome variables at next visit

The ability of the time series models to predict new concentration values in any upcoming clinical visit was evaluated using the Spearman correlation (between the actual values from the CRF and the predicted ones) in the LOPO set-up. The correlation values for both, models trained with the CRF model variables and models

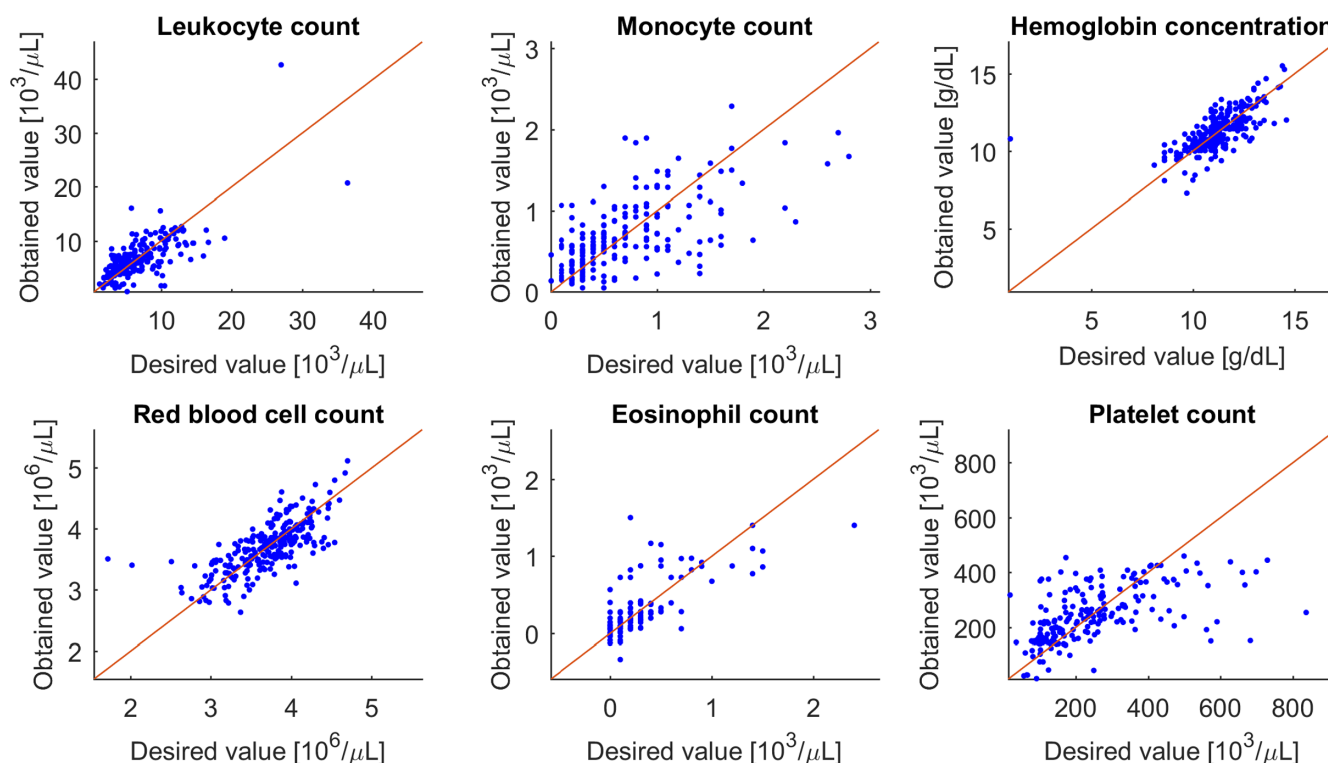
trained with the CRF plus TPMS model variables, are listed in Table 2.

Reasonably good correlation coefficients were obtained in predicting the concentration values for two of the outcome variables: red blood cells and hemoglobin. In this study, the inclusion of variables obtained from the TPMS models did not have a significant impact in the accuracy of the predictions.

As we are aiming at providing an individualized prediction, we computed the correlations for each patient separately and averaging all the correlation values obtained. This led to weaker correlation values for the six outcome variables (Table 3) for both sets of models, CRF and CRF + TPMS, thus making these models not eligible for the prediction of expected concentration values in the following visits.

### Ability to predict trends of outcome variables at next visit

To measure the ability of the models to predict increasing or decreasing trends for the six selected outcomes compared to the previous visit's values, prediction scores expressed as balanced accuracies were calculated between the predicted and observed trend, using the LOPO setup



**FIGURE 2** Predicted versus observed plot for the six outcome variables. The observed values (desired value) correspond to the CRF actual value. Predicted values (obtained value) correspond to the model trained with CRF model variables in the LOPO setup. The red line corresponds to the identity /i.e., where the points would ideally lie). CRF, Case Report Form; LOPO, leave-one-patient-out.

**TABLE 2** Concentration value prediction, correlation computed taking all patients together

Outcome variable	Spearman Correlation coefficient	
	CRF	CRF + TPMS
Leukocytes	0.71	0.71
Monocytes	0.64	0.60
Hemoglobin	0.76	0.77
Red blood cells	0.76	0.78
Eosinophils	0.72	0.71
Platelets	0.68	0.67

Note: Spearman correlation between observed and predicted concentration values trained with LOPO setup. CRF: models trained with the CRF variables alone. CRF + TPMS: models trained with the CRF and TPMS derived variables.

Abbreviations: CRF, Case Report Form; LOPO, leave-one-patient-out; TPMS, Therapeutic Performance Mapping System.

**TABLE 3** Concentration value prediction, correlation calculated per patient

Outcome variable	Spearman correlation coefficient (computed per patient)	
	CRF	CRF + TPMS
Leukocytes	0.37	0.36
Monocytes	0.47	0.44
Hemoglobin	0.45	0.51
Red blood cells	0.45	0.52
Eosinophils	0.39	0.30
Platelets	0.40	0.35

Note: Spearman correlation between observed and predicted concentration values considering patients individually (i.e., the correlation is computed per patient and the average among patients is given). Models trained with LOPO set up. CRF: models trained with the CRF variables alone. CRF + TPMS: models trained with the CRF and TPMS derived variables.

Abbreviations: CRF, Case Report Form; LOPO, leave-one-patient-out; TPMS, Therapeutic Performance Mapping System.

for cross-validation. The binary outcome for this setup was obtained from the regression outcomes of the models (in the LOPO setup). The transformation to a binary outcome was performed by labeling the predicted regression values that are higher (respectively lower) than the true outcome at the previous visit as one (respectively 0). This binary outcome was then compared to its true corresponding outcome (i.e., 1/0 if the true value at the objective visit is higher/lower than the true value at the previous visit) to compute the prediction scores. Note that the visits for which there is no increase or decrease compared to the previous visits were ignored in the computation

of this score. The prediction scores reflect the ability of the model to classify an increase versus a decrease of an outcome variable value compared to its value at the last known visit. As seen in [Table 4](#), the two outcome variables that performed best when predicting concentration values ([Table 2](#)), hemoglobin concentration, and red blood cells count, were the worst when predicting trends. Good prediction scores were obtained for the rest of the outcome variables, being the prediction of eosinophil concentration trend the best.

## Relevant clinical and molecular variables in the prediction models

The variables selected by the models when predicting trends with respect to the median value of a patient outcome can be considered as the factors that had the greatest influence on the prediction. The model for each outcome variable may involve several trees, the final prediction is the arithmetic mean of all of the models' single-tree predictions. A summary of the variables that the models have selected for predicting the different outcomes is shown in [Table 5](#). The relevant variables per outcome are shown grouped per tree and with both clinical features from the CRF and proteins obtained from the TPMS models. All model trees detailing the relations between the variables are provided in the [Data S1](#).

Overall, we see a strong association with the time variable, followed by neutropenia, low concentration of neutrophils in blood. These two variables are present in almost all the models, the possible implications are discussed in the Discussion section. The potential role of other factors contributing to the prediction of only one of the outcome variables are discussed in this section.

In most instances, it is difficult to establish a linear correlation between the outcome variable and any of the variables selected by the model, as the relation established by the models depends on different cutoff thresholds combined with other variables. To get a full picture of the complex dependencies between the variables established by the models to reach a prediction, one needs to consult the regression trees from the models (available in [Data S1](#)).

## Eosinophils

The prediction of eosinophil trends was the one with the best score, with a balanced accuracy of 0.84 for models with only CRF model variables and 0.86 including TPMS model variables. The prediction appears to be only dependent on the time span between visits and the neutropenia

**TABLE 4** Prediction scores (balanced accuracy) to measure the ability to predict increase versus decrease compared to the previous visit's outcome's value

Outcome measures	Prediction score (Balanced accuracy)	
	CRF	CRF + TPMS
Leukocytes	0.77	0.76
Monocytes	0.77	0.77
Hemoglobin	0.61	0.68
Red blood cells	0.61	0.67
Eosinophils	0.84	0.86
Platelets	0.77	0.76

Note: CRF: models trained with the CRF variables alone. CRF + TPMS: models trained with the CRF and TPMS derived variables.

Abbreviations: CRF, Case Report Form; TPMS, Therapeutic Performance Mapping System.

values at different previous visits, in both sets of CRF and CRF + TPMS models.

## Leukocytes

The models predicting leukocyte trends gave a balanced accuracy of 0.77 for models with only CRF model variables and 0.76, including TPMS model variables. Again, time and neutropenia were part of all regression trees. Models built using only CRF data combined time and neutropenia with the use of either fentanyl in the penultimate visit or naloxone in the last visit. Both drugs target the opioid receptors but with opposite effects, naloxone is given to reverse a fentanyl overdose. Moreover, neuropathy, a side effect of chemotherapy, is one of the contributing factors in one of the prediction models, this condition is frequently treated with opioid drugs such as fentanyl. Thus, it is likely that the three different variables come from the same effect. Molecular interrelations of neuronal and hematopoietic signaling<sup>19,20</sup> have been described and could account for the correlation identified by this model.

The models predicting leukocyte trends from CRF and TPMS data combined time and neutropenia with five different proteins. All the proteins appearing in the models can be related at different levels with hematopoiesis regulation.<sup>21–25</sup>

## Monocytes

Once more, all models predicting monocyte trends use time and neutropenia variables at last visit with a balanced

accuracy of 0.77. In the models based on CRF data only, sodium values at the last or last two visits appear also to be relevant, together with the previous two factors. Other variables that combine with those already named are creatinine and metastasis.

Previous studies<sup>26</sup> showed a strong positive association between salt-intake levels and monocyte numbers. Sodium activates human monocytes to differentiate and promote local inflammation. Although this activation of the immune system may be deleterious for healthy subjects, recent studies show that local concentration of sodium chloride in tumor tissue inhibits tumor growth by activating immune surveillance.<sup>27</sup> This could also explain the relevance of the metastasis variable together with sodium in the prediction of monocyte trends.

Both creatinine and sodium are markers of kidney function. Increased serum creatinine levels and low levels of blood sodium can result in decreased kidney function. Furthermore, peripheral neutrophil count and monocyte count has been associated with renal progression.<sup>28,29</sup> This could provide an explanation for the inclusion of kidney-related factors in the prediction of monocyte trends.

At the molecular level, the CYP7B1 protein was included in the monocyte prediction models together with time and neutropenia. This enzyme catalyzes the first reaction in the cholesterol catabolic pathway, controls the levels of intracellular regulatory oxysterols. Oxysterols are not only regulators of cholesterol homeostasis but also have important roles in the control of immune responses, they have a stimulatory effect driving the differentiation of monocytes.<sup>30</sup>

## Platelets

Platelet prediction trends gave balanced accuracies of 0.77 and 0.76 for CRF and CRF + TPMS models and also included time span between visits in all the trees. Neutropenia was not as universal as in the previously described models but was also present. The models that did not include neutropenia included lymphocyte count. The links between hematological cell counts can be easily explained by the common origin.

Those variables were combined with different AEs, mucositis, diarrhea, and arthritis. The contribution of these factors in platelet prediction are less obvious. Mucositis and diarrhea are common side effects in chemotherapy treatments, in both conditions, low platelet counts could pose a risk of internal bleeding, but no causal relationships have been described. Arthritis as a side effect of chemotherapy is rarer but a pathogenic involvement of

**TABLE 5** Variables selected by the regression models predicting increasing or decreasing trends

Outcome	CRF		CRF + TPMS	
	Variable	Time of measurement	Variable	Time of measurement
Eosinophils	Time	Last & antepenultimate visits (1, 3)	Time	Last visit (1)
	Neutropenia	Last two visits (1, 2)	Neutropenia	Last visit (1)
Leucocytes	Time	Last two visits (1, 2)	Time	Last two visits (1, 2)
	Neutropenia	Last visit (1)	Neutropenia	Last visit (1)
	Time	Last two visits (1, 2)	ROCK1	Last two visits (1, 2)
	Neutropenia	Last visit (1)	Time	Last two visits (1, 2)
	Neuropathy	Penultimate visit (2)	Neutropenia	Last visit (1)
	Time	Last two visits (1, 2)	HSPA5	Penultimate visit (2)
	Neutropenia	Last visit (1)	Time	Last two visits (1, 2)
	DB00813 (Fentanyl)	Penultimate visit (2)	Neutropenia	Last visit (1)
	Time	Last two visits (1, 2)	TBK1	Last visit (1)
	Neutropenia	Last visit (1)	Time	Last two visits (1, 2)
	DB01183 (Naloxone)	Last visit (1)	Neutropenia	Last visit (1)
			IL23R	Penultimate visit (2)
			Time	Last two visits (1, 2)
			Neutropenia	Last visit (1)
			CCR1	Penultimate visit (2)
Monocytes	Time	Last visit (1)	Time	Last visit (1)
	Neutropenia	Last visit (1)	Neutropenia	Last visit (1)
	Sodium	Last two visits (1, 2)	CYP7B1	Last visit (1)
	Creatinine	Last two visits (1, 2)		
	Time	Last two visits (1, 2)		
	Neutropenia	Last visit (1)		
	Sodium	Penultimate visit (2)		
	Metastasis	Penultimate visit (2)		
Platelets	Time	Last three visits (1–3)	Time	Last, antepenultimate (1, 3)
	Neutropenia	Penultimate visit (2)	POLD3	Last, antepenultimate (1, 3)
	Mucositis	Last visit (1)	Time	Last, antepenultimate (1, 3)
	Time	Last three visits (1–3)	POLE3	Last, antepenultimate (1, 3)
	Neutropenia	Last visit (1)	Time	Last, antepenultimate (1, 3)
	Diarrhea	Last two visits (1, 2)	POLE	Last, antepenultimate (1, 3)
	Time	Last three visits (1–3)	Time	Last, antepenultimate (1, 3)
	Lymphocytes	Last three visits (1–3)	POLD4	Last, antepenultimate (1, 3)
	Arthritis	Antepenultimate (3)	Time	Last, antepenultimate (1, 3)
	Time	Last three visits (1–3)	POLE4	Last, antepenultimate (1, 3)
	Lymphocytes	Last, antepenultimate (1, 3)		
	Mucositis	Last visit (1)		
Red blood cells	Red blood cells	Last visit (1)	ERN1	Last visit (1)
			EIF2AK3	Last visit (1)
Hemoglobin	Time	Last visit (1)	NOTCH4	Last visit (1)
	Hemoglobin	Last visit (1)	TRPC6	Last visit (1)
			NOTCH4	Last visit (1)
			RNF111	Last visit (1)

*Note:* CRF: models trained with the CRF variables alone. CRF + TPMS: models trained with the CRF and TPMS derived variables. The variables are listed per groups (with or without gray filling) to show the tree grouping for those cases where the variables differ in the different trees.

Abbreviations: CRF, Case Report Form; TPMS, Therapeutic Performance Mapping System.

platelets in joint inflammation has been described, pointing to the existence of crosstalk between the coagulation and inflammation systems.<sup>31</sup>

The prediction models with TPMS data combine the time factor with five proteins, all related with DNA replication and repair. They may have direct relationship with the effect of gemcitabine treatment that disrupts normal DNA synthesis, as sometimes DNA repair processes are activated contributing to resistance. But no obvious relationship with platelet count can be deduced. The prediction models including these proteins are not as good as the models with only the CRF model variables.

## Hemoglobin and red blood cells

The models predicting hemoglobin concentration and red blood cell count trends at next visit did not reach as good accuracy values as those obtained for the other outcome variables, with a balanced accuracy of 0.61 and 0.68–0.67 for CRF and CRF + TPMS models, respectively. Prediction models built with only clinical data based the prediction on the concentration from the previous visit of the variable itself. Hemoglobin prediction also took into consideration the time span from the last visit.

The prediction models built with clinical and TPMS data based the prediction on the variations of proteins streaming from TPMS only and no clinical variables were included. The two proteins used by the models to predict red blood cell count trends, ERN1 and EIF2AK3, are both related to unfolded protein stress response, a mechanism activated to improve blood cell counts after irradiation.<sup>32</sup>

Notch4 appears in all the models predicting hemoglobin trends. Notch signaling has a role in regulation of adult steady-state bone marrow myelopoiesis, including erythropoiesis. In fact, targeting Notch signaling has been proposed as a new therapeutic approach to mitigate chemotherapy-induced injury.<sup>33</sup>

Transient receptor potential canonical channel 6 (TRPC6), included in the hemoglobin model, participates in cation leak and Ca(2+)-induced suicidal death in red blood cells<sup>34</sup> and can be thus related to detection of hemoglobin in blood.

A more distant relationship between RNF111 and hemoglobin count can be established. RNF11 is an E3 ubiquitin ligase part of the TGF- $\beta$  signaling pathway, activates SMAD-dependent transcription in response to TGF- $\beta$ . This is a pleiotropic pathway with a role in erythroid differentiation through Smad4-mediated growth inhibition in response to TGF- $\beta$ .<sup>34</sup>

## Tool for aiding clinical decision making

As reading and interpreting the regression trees is not intuitive, we have built a user-friendly interface, deployed as a webapp accessible at <http://sicpac.anaxomics.com:81>, to query the models and facilitate usage and validation in scientific and clinical settings.

The web-application allows users to predict the trend of the outcome variables (eosinophils, leukocytes, monocytes, platelets, red blood cells, and hemoglobin) at the next visit, based on 18 clinical variables from up to three previous visits and the median value for the variable from all previous visits.

The following variables are provided as yes or no input: fentanyl prescription, naloxone prescription, neuropathy, neutropenia, metastasis, arthritis, diarrhea, and mucositis. The following variables are input as concentration values: creatinine, sodium, lymphocytes, red blood cells, hemoglobin, leukocytes, monocytes, eosinophils, and platelets.

## DISCUSSION

We explored the ability of machine-learning models with the support of systems-biology inputs to provide a clinical decision-support tool that could aid the evaluation of patients with pancreatic cancer progression in response to treatment, focusing on the prediction of the evolution of routinely recorded clinical variables at the next visit.

In selecting the modeling strategy, several factors had to be accounted for. The fact that we are dealing with real-world data with missing values, with mixing of categorical and continuous variables, and a reduced number of patients, posed the first limitations on the strategies that could be applied. The patterns in the routinely collected medical data are complex and we wanted a model that could cope with nonlinear interactions between associated factors and the outcome. We selected decision trees as the core for the mathematical models as these complied with the previously mentioned limitations and requirements, with the added value of the model being, to some extent, easily and directly interpretable.

The selection of the outcome variables was based on the information value for the clinician but also on the variation range in the population under study. Thus, variables like metastasis stage or the Karnofsky or ECOG scales, with a limited variation among the patients under study, were not suitable. Blood counts offered both, variation among patients and information value as they may indicate comorbid conditions, the extent of disease, or individual response to hematological toxicity of chemotherapy.

Thus, a model capable of predicting the evolution of hematological parameters would be of great help for the clinical management of patients, for example, to adjust therapy dosage.

When predicting the concentration values of the outcome variables, the models achieved a correlation value between true and predicted outcome of around 0.74 when considering all patients together. However, the models are aimed at predicting individual patient values and the correlation measure should thus be calculated also at this level. The correlation dropped to an average correlation of around 0.42 when considering the patients separately; thus, neither model (CRF or CRF + TPMS) was eligible for predicting expected concentration values at the next visit when aiming at individualized predictions. Although the models were not able to predict the value of the outcome variables on an individual basis, which was our most ambitious aim, they were still able to predict trends (increase or decrease) of the outcome variables at this individual level, especially for eosinophils, leukocytes, monocytes, and platelets, with a mean prediction score (balanced accuracy) of 0.79. When grouping outcome variables by the accuracy values of the prediction models, they tend to cluster by hematopoiesis lineage. As eosinophils and monocytes are part of the leukocyte count, similar behavior can be expected. Red blood cells are packed with hemoglobin; thus, their concentration trends are also more likely correlated.

When evaluating the factors (variables) that the models selected to predict the outcome variables, only the models' predicting trends for hemoglobin or red blood cells used the value from the previous visit of the same variable being predicted. Taking into consideration that the median time between visits was of 7 days, this may be a reflection of repopulation kinetics in the marrow, with hemoglobin and erythrocyte having longer replenishment rates (up to 84 days in healthy individuals) compared to leukocytes (7 days in healthy individuals). In fact, the time span between visits was the most common factor included in almost all the models, probably reflecting differential rates of recovery after chemotherapy treatment for the different cell types and the health status of the patient.

The second most common factor in the prediction models was neutropenia, a frequent AE deriving from chemotherapy treatment. The fact that lower neutrophil counts may reflect on other blood cell counts is not surprising and could be indicative of bone marrow depletion. But the exact relationship between the variables in our models is complex and does not necessarily need to be linear, thus we cannot establish the type of relationship at play between the variables. On the other hand, other frequent blood related AEs present in the study, like lymphopenia,

thrombocytopenia, or anemia, did not provide prediction power to the models. Thus, we think that neutrophil count is especially relevant to assess patient status. In fact, neutropenia appears to be more than just an AE and there are many studies highlighting its potential as a surrogate marker of response and/or survival in patients treated with cytotoxic regimens.<sup>35</sup>

The inclusion of molecular variables together with the clinical variables (CRF + TPMS models) had a marginal effect on the prediction scores compared to the models with only clinical variables (CRF models). It is possible that, because the TPMS models are built using CRF information, they might somehow correlate with a combination of them and not increase performance. Nevertheless, the inclusion of TPMS derived variables was meant to incorporate molecular-level information, potentially revealing some of the pathways underlying the forecasted progression of the outcome variable, which can help interpretation and guide further research. But they are not relevant for clinical application of the prediction models because the TPMS model would not be available in the hospital-visit setup.

The proteins selected by the models as relevant for predicting next-visit outcome variables were all different between the different models, but a literature search on already described relationships with the outcome variables they are predicting showed some common paths. The individual protein links have been discussed along the Results section. No molecular factors were selected for eosinophil prediction. The proteins involved in the prediction of leukocytes, monocytes, red blood cells, and hemoglobin trends have all a relation with hematopoiesis regulation. The models selected proteins related to DNA replication only in the case of platelets, probably in connection to the treatment's mode of action.

In conclusion, the modeling strategy applied in this study could determine the tendency of hematological parameters at next visit based on standardly collected clinical parameters from patients with pancreatic cancer treated with a gemcitabine/nab-paclitaxel regime as first-line palliative chemotherapy. It also provided clinical and molecular understanding of the measured factors that contribute the most to the prediction of the outcome variables' evolution.

The relatively low performance predicting the exact value of the outcome variables limits the application of this methodology to the prediction of trends, at least when the amount of data available is low, as was the case here. The reduced number of patients with pancreatic cancer from a single center and with the same chemotherapy does also limit the universality of the prediction. Nevertheless, this study serves as a proof of concept for the application of machine-learning tools to the prediction of next-visit

outcomes in real-world settings, highlighting its usefulness even when the data set might be considered small.

## AUTHOR CONTRIBUTIONS

V.J., J.M.G.I., E.R., B.O., J.F., X.D., and J.M.M. wrote the manuscript. E.R., J.M.M., B.O., X.D., and R.M. designed the research. E.R. and R.M. performed the research. V.J., J.F., P.M.F., J.M.G.I., and J.M.M. analyzed the data. V.J. and P.M.F. contributed new analytical tools.

## FUNDING INFORMATION

V.J. is part of a project (COSMIC) that has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 765158. P.M.F. receives funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 860303. J.M.G.I. receives funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 859962.


## CONFLICT OF INTEREST STATEMENT

V.J. was an employee of Anaxomics Biotech S.L. at the time of the study. P.M.F., J.M.G.I., and J.F. are employees of Anaxomics Biotech S.L. J.M.M. is an employee and co-founder of Anaxomics Biotech S.L. B.O. and X.D. are scientific advisors for Anaxomics Biotech but do not perceive any financial support for these services. All other authors declared no competing interests for this work.

## ORCID

Valentin Junet  <https://orcid.org/0000-0002-6138-0612>

Pedro Matos-Filipe  <https://orcid.org/0000-0003-0492-0683>

Juan Manuel García-Illarramendi  <https://orcid.org/0000-0003-3368-814X>

Baldo Oliva  <https://orcid.org/0000-0003-0702-0250>

Judith Farrés  <https://orcid.org/0000-0002-0958-0510>

Xavier Daura  <https://orcid.org/0000-0001-9235-6730>

## REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin.* 2018;68:7-30.
2. Garrido-Laguna I, Hidalgo M. Pancreatic cancer: from state-of-the-art treatments to promising novel therapies. *Nat Rev Clin Oncol.* 2015;12:319-334.
3. Von Hoff DD, Ervin T, Arena FP. Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N Engl J Med.* 2013;369:1691-1703.
4. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N Engl J Med.* 2011;364:1817-1825.
5. Ducreux M, Seufferlein T, van Laethem JL, et al. Systemic treatment of pancreatic cancer revisited. *Semin Oncol.* 2019;46:28-38.
6. Sohal DPS, Kennedy EB, Cinar P, et al. Metastatic pancreatic cancer: ASCO guideline update. *J Clin Oncol.* 2020;38:3217-3230.
7. Enderling H, Alfonso JCL, Moros E, Caudell JJ, Harrison LB. Integrating mathematical modeling into the roadmap for personalized adaptive radiation therapy. *Trends Cancer.* 2019;5:467-474.
8. Radiology E.S.o. What the radiologist should know about artificial intelligence – an ESR white paper. *Insights into Imaging.* 2019;10:1-8.
9. Li D, Pehrson LM, Lauridsen CA, et al. The added effect of artificial intelligence on Physicians' performance in detecting thoracic pathologies on CT and chest X-ray: a systematic review. *Diagnostics.* 2021;11: 2206-2222.
10. Simon G, DiNardo CD, Takahashi K, et al. Applying artificial intelligence to address the knowledge gaps in cancer care. *Oncologist.* 2019;24:772-782.
11. Nedungadi P, Iyer A, Gutjahr G, Bhaskar J, Pillai AB. Data-driven methods for advancing precision oncology. *Curr Pharmacol Rep.* 2018;4:145-156.
12. Reardon B, Moore ND, Moore NS, et al. Integrating molecular profiles into clinical frameworks through the molecular oncology almanac to prospectively guide precision oncology. *Nat Can.* 2021;2:1102-1112.
13. Jorba G, Aguirre-Plans J, Junet V, et al. In-silico simulated prototype-patients using TPMS technology to study a potential adverse effect of sacubitril and valsartan. *PLoS One.* 2020;15:e0228926.
14. Morales R, Serrano R, Sardón T Functional, structural and contextual analysis of a variant of uncertain clinical significance in BRCA1: c.5434c>G (p.pro1812ala). *J Cancer Genet Biomark.* 2017;1:1.
15. Herreros-Villanueva M, Pére-Palacios R, Castillo S. Biological relationships between miRNAs used for colorectal cancer screening. *J Mol Biomark Diag.* 2018;9:398.
16. Carcereny E, Fernández-Nistal A, López A, et al. Head to head evaluation of second generation ALK inhibitors brigatinib and alectinib as first-line treatment for ALK+ NSCLC using an in silico systems biology-based approach. *Oncotarget.* 2021;12:316-332.
17. Gutierrez-Casares JR, Quintero J, Jorba G. Methods to develop an in silico clinical trial: computational head-to-head comparison of Lisdexamfetamine and methylphenidate. *Front Psych.* 2021;12:741170.
18. Django Software Foundation. Django. 2019. Available at: <https://djangoproject.com>. Accessed May 20, 2021.
19. Hu N, Yu T, Chen J, Zheng S, Yan H, Duan J. Oxycodone stimulates normal and malignant hematopoietic progenitors via opioid-receptor-independent- $\beta$ -catenin activation. *Biochem Biophys Res Commun.* 2020;533:1457-1463.
20. Hanoun M, Maryanovich M, Arnal-Estapé A, Frenette PS. Neural regulation of hematopoiesis, inflammation, and cancer. *Neuron.* 2015;86:360-373.
21. Kapur R, Shi J, Ghosh J, et al. ROCK1 Via LIM kinase regulates growth, maturation and Actin based functions in mast cells. *Oncotarget.* 2016;7:16936-16947.
22. Leveau C, Gajardo T, el-Daher MT, et al. Ttc7a regulates hematopoietic stem cell functions while controlling the stress-induced response. *Haematologica.* 2020;105:59-70.
23. Louis C, Burns C, Wicks I. TANK-binding kinase 1-dependent responses in health and autoimmunity. *Front Immunol.* 2018;9:434.

24. Márquez S, Fernández JJ, Terán-Cabanillas E, et al. Endoplasmic reticulum stress sensor IRE1 $\alpha$  enhances IL-23 expression by human dendritic cells. *Front Immunol*. 2017;8:639.
25. Zilio S, Biciato S, Weed D, Serafini P. CCR1 and CCR5 mediate cancer-induced myelopoiesis and differentiation of myeloid cells in the tumor. *J Immunother Cancer*. 2022;10:e003131.
26. Yi B, Titze J, Rykova M, et al. Effects of dietary salt levels on monocytic cells and immune responses in healthy human subjects: a longitudinal study. *Transl Res*. 2015;166:103-110.
27. He W, Xu J, Mu R, et al. High-salt diet inhibits tumour growth in mice via regulating myeloid-derived suppressor cell differentiation. *Nat Commun*. 2020;11:1732.
28. Bowe B, Xie Y, Xian H, Li T, Al-Aly Z. Association between monocyte count and risk of incident CKD and progression to ESRD. *Clin J Am Soc Nephrol*. 2017;12:603-613.
29. Yen CH, Wu IW, Lee CC, et al. The prognostic value of peripheral total and differential leukocyte count in renal progression: a community-based study. *PLoS One*. 2021;16:e0258210.
30. Son Y, Kim SM, Lee SA, Eo SK, Kim K. Oxysterols induce transition of monocytic cells to phenotypically mature dendritic cell-like cells. *Biochem Biophys Res Commun*. 2013;438:161-168.
31. Targońska-Stępnia B, Grzechnik K, Zwolak R. The relationship between platelet indices and ultrasound, clinical, laboratory parameters of disease activity in patients with rheumatoid arthritis. *J Clin Med*. 2021;10: 5259-5270.
32. Luchsinger LL. Hormetic endoplasmic reticulum stress in hematopoietic stem cells. *Curr Opin Hematol*. 2021;28:417-423.
33. Chen J, Dong Y, Peng J, et al. Notch signaling mitigates chemotherapy toxicity by accelerating hematopoietic stem cells proliferation via c-Myc. *Am J Transl Res*. 2020;12:6723-6739.
34. Foller M, Kasinathan RS, Koka S. TRPC6 contributes to the Ca(2+) leak of human erythrocytes. *Cell Physiol Biochem*. 2008;21:183-192.
35. Kasi PM, Grothey A. Chemotherapy-induced neutropenia as a prognostic and predictive marker of outcomes in solid-tumor patients. *Drugs*. 2018;78:737-745.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Junet V, Matos-Filipe P, García-Illarramendi JM, et al. A decision support system based on artificial intelligence and systems biology for the simulation of pancreatic cancer patient status. *CPT Pharmacometrics Syst Pharmacol*. 2023;12:916-928. doi:[10.1002/psp4.12961](https://doi.org/10.1002/psp4.12961)